Empirical Analysis of Sequence Length and Training Epochs for AraBERT in Arabic News Classification

Yasir Hadi Farhan

Department of Artificial Intelligence, College of Information Technology,

University of Fallujah, Iraq

Abstract

Arabic news classification has become an important downstream task for Arabic natural language processing (NLP), especially as news portals continue to generate large volumes of category-rich content. Although Arabic transformer models have improved performance on many classification tasks, practical deployment still depends heavily on seemingly simple hyperparameters such as the maximum input length and the number of fine-tuning epochs. This paper presents an empirical study of these two factors using AraBERTv2 on a balanced subset of the MAAD Arabic news corpus. We evaluate three sequence lengths (64, 128, and 256 tokens) and three epoch settings (1, 2, and 3) under the same tokenizer, optimizer family, and batch size. The results show that a sequence length of 128 tokens yields the best performance in the length experiment, with 97.18% accuracy and 97.17% macro-F1, while the best epoch setting is three epochs, reaching 97.44% accuracy and 97.43% macro-F1. The results indicate that moderate sequence lengths are sufficient for this task, and that controlled additional training improves performance without obvious overfitting within the tested range. The study provides a concise, reproducible set of guidelines for configuring AraBERT for Arabic news classification.

Keywords: Arabic NLP; AraBERT; Arabic news classification; sequence length; training epochs; hyperparameter analysis.

1. Introduction

One of the most common tasks in Arabic NLP is Arabic text classification since it facilitates document indexing, content filtering, recommendation, search, and media analytics. In news fields, proper topic classification is of great essence since news pieces tend to be short to medium in length, topically congested and high volume. Simultaneously, Arabic poses sustained representation learning troubles, such as rich morphology, variation of spelling, and sparsity of tokens, and coexisting Modern Standard Arabic and dialectal forms [1, 2].

The latest trend in Arabic large language models and pretrained transformer encoders has made a significant contribution to the quality of Arabic text representations. Models and benchmarks published after 2023 Arabic-centered models and benchmarks models have indicated the fast pace of the field through larger Arabic-centered foundation models, more competitive encoder benchmarks, and dialect-specific pretrained models [3-7]. These experiments establish that pretraining quality, domain diversity, and even evaluation design are all important but they also demonstrate that downstream performance is also sensitive to fine-tuning choices [4, 8].

The configuration of hyperparameters is critical in classification. The parameter with the most practical and usage is the maximum sequence length and the number of training epochs. The length of the sequence determines the extent of information of the input text that the model is exposed to whereas the number of epochs determines the extent to which the model is able to adjust to the task at hand. In the classification of Arabic texts, sequences that are too short risk the loss of valuable context, whereas too long sequences can add noise and raise the cost of computation. Similarly, there are two extreme cases: underfitting of too few epochs and overfitting of too many epochs [9-12].

Although the Arabic transformer studies are increasing, relatively few of them isolate the contribution of these basic settings in a localized Arabic news classification situation. In this paper, the gap is filled in by a concise empirical investigation based on AraBERTv2 and a balanced sample of the MAAD Arabic news data. The focus is deliberately small and practical: the study does not suggest any new architecture but inquires about the most promising environments to a realistic Arabic news classification pipeline.

There are fourfold contribution of this paper. First, it gives a reproducible study analysis of sequence length of AarBERT-based Arabic news categorization. Second, it assesses the impact of training epochs in fixed circumstances. Third, it reports actual measured outputs with an Arabic balanced news corpus. Fourth, it provides configuration advice which can help future small-scale Arabic NLP research and local conference work.

2. Related Work

The study of Arabic transformers has gone very fast over the past few years. Jais and Jais-chat presented the possibility of big Arabic-oriented generative models and the necessity of more robust Arabic resources and assessments [3]. On the encoder level, data scale in Arabic pretraining, it has been demonstrated that retraining or scaling high-quality Arabic corpora could be improved significantly on standard benchmarks [4]. Swan and ArabicMTEB further expanded Arabic evaluation by providing an Arabic centric and dialect sensitive suite of benchmarks to embeddings and retrieval-style evaluations [5].

Dialect-based pretrained models have gained greater significance in addition to general Arabic models. Each of the three studies (SaudiBERT, EgyBERT, and AlcLaM) demonstrates that domain and dialect specialization can have a significant material effect on downstream performance in classification, particularly with low-resource or dialect-laden environments [6, 7, 13]. The importance of these results is that Arabic news is not always pure uniform Modern Standard Arabic; topic, source, and style of writing may influence the effective transfer to downstream classification when a pretrained model is used.

Task specific studies are also several recent studies which confirm the usefulness of transformer-based Arabic classification. A cross-dialect banking intent classification system that relies on transformers delivered good performance when using the AraBERTv2 model and, compared to traditional machine learning and simpler deep learning baselines, it performs better [14]. The Arabic Text Classification on Functional Text Dimensions offered a new Arabic text classification corpus and benchmark with excellent performance on the transformer-based methods compared to more conventional baselines [15]. The detection of Arabic hate-speech, classification of long-tweets, and dialect sentiment tasks based on pretrained language models and data augmentation have been studied in other work, further proving the applicability of transformer fine-tuning in Arabic classification [9, 16, 17].

The other line of related research is related to optimization and configuration. Arabic BERT models Studies of the impact of pretraining corpora and model selection have been conducted [8], as has more general research on text classification in right-to-left languages which demonstrates that the details of sequence processing can be combined with the value of Arabic representations [10]. Status Studies on Arabic classification in recent times have still indicated that tuning choices, including maximum length, training steps, and checkpoint choice can observeably impact performance, despite the underlying backbone model is held constant [11, 12, 18].

It is based on this literature that the current study will be devoted to two directly actionable settings, that is, sequence length and epoch count. The decision is inspired by experience. The following are some of the first parameters that are tuned by researchers that conduct small- or medium-scale Arabic classification experiments, but with few reproducible results published about the impacts of these parameters on Arabic news classification using AraBERT.

3. Methodology

3.1 Dataset

The experiments were performed on the MAAD (Multi-Label Arabic Articles Dataset), which is an Arabic news corpus that was gathered in the internet news sources. This paper has used a balanced sample of six categories of news in the creation of a balanced subset. The classes had a contribution of about 2900 instances each, which made a balanced dataset that could be fairly compared with different settings of hyperparameters.

3.2 Preprocessing

The preprocessing pipeline eliminated the URLs, minimized the number of repeated whitespaces, and filtered out uninformative symbols. The last experiments involved the identical clean-up data in all experiments. The objective of the preprocessing phase was not to aggressively normalize the news text but to stably prepare the news text in such a way that the comparison was based on differences in hyperparameters and not variation in data-quality.

3.3 Model

AarBERTv2 was employed as the backbone in the experiments. AraBERTv2 is a pretrained Arabic-oriented transformer encoder that is extensively applied in Arabic classification. The head of the task was introduced based on a common single label classification layer located on top of the transformer encoder.

3.4 Experimental design

Two experiments were conducted. In the former, the highest sequence length was varied at 64, 128 and 256 tokens as the rest of the settings were fixed. In the second, the number of epochs was changed between 1, 2 and 3 keeping the rest of the configuration constant. The training batch size and learning rate used in both experiments was 8 and 2e-5 respectively. As evaluation metrics, accuracy and macro-F1 were considered.

3.5 Evaluation metrics

The reason why accuracy was chosen is due to the fact that the dataset is balanced and the task is single-label multi-class classification. There was also the report of Macro-F1 to make sure that every single class had an equal share in the evaluation in spite of any local performance variation. Accuracy alone does not give a complete picture as compared to reporting both metrics.

3.6 Implementation details

Experiments were executed all in Google Colab in the Hugging Face Transformers library. The tokenizer and encoder backbone were held constant throughout (experimental settings) so that only the appropriate hyperparameter varied in an experiment. The design will minimize confounding variation and simplify the interpretation of the results. All experimental branches experimented had the same train/test set-up and the same evaluation routine used at the conclusion of each run.

4. Experimental Results

4.1 Sequence length analysis

Results of sequence-length experiment are given in table 1. The highest score was 128 tokens where the accuracy and macro-F1 were 97.18 and 97.17 respectively. Even though the performance of 256 tokens was still good, it was not higher than in the case of 128-token arrangement. This implies that, in the case of the samples of news utilized in this research, the key discriminative information is greatly obtained within a moderate sequence budget.

Table 1. Effect of maximum sequence length on AraBERTv2 performance.

Sequence Length	**Accuracy**	**Macro-F1**
64	0.9626	0.9624
128	0.9718	0.9717
256	0.9695	0.9694

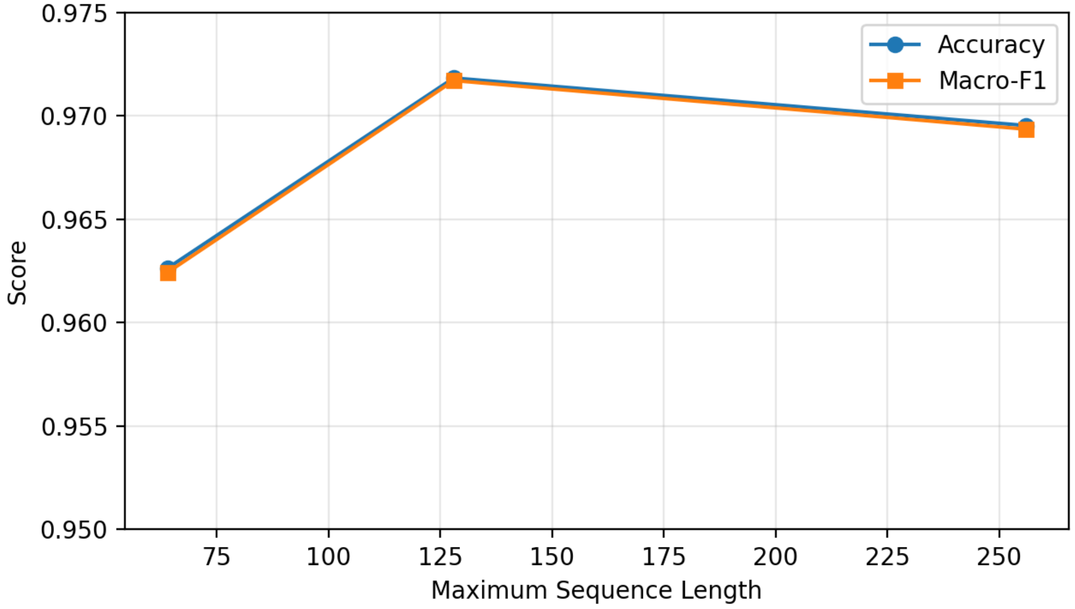

Figure 1. Accuracy and macro-F1 across sequence-length settings.

The same trend can be visualized in Figure 1. The 64-token environment already provides good competitive results, which means that the classification indicator is quite strong on the news sample. But there is a definite improvement when the number of tokens is increased by 64 to 128. The minor decrease in the number of tokens (128 to 256) indicates that it is not necessarily good to provide the model with more context. It is possible that a reason is that longer sequences provide background information, which is not needed to predict the category, and may overpower the most significant features.

4.2 Epoch analysis

The epoch experiment is shown in table 2. The outcomes become better in epoch to epoch. Three epochs produced the highest accuracy of 97.44% and macro-F1 of 97.43% and was the best configuration. In the range being tested, no harmful effects of overfitting were noticed; the model further still had the advantage of further supervised learning the task.

Table 2. Effect of epoch count on AraBERTv2 performance.

Epochs	Accuracy	Macro-F1
1	0.9601	0.9599
2	0.9716	0.9714
3	0.9744	0.9743

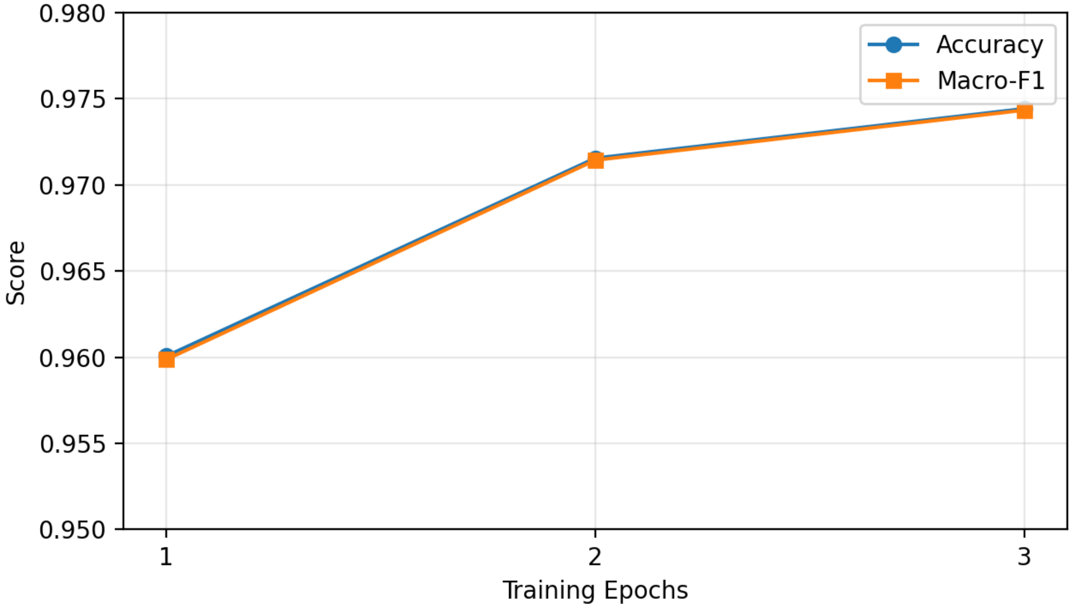

Figure 2. Accuracy and macro-F1 across epoch settings.

According to Figure 2, both measures are increasing in a monotonic way with the increasing number of epochs(1-3). This action gives credence to the fact that the model could still have evolved the first and second epochs. The results are also useful to researchers who are constrained by time or computational resources since they provide a quantification of the trade-off: one epoch already achieves approximately 96% accuracy, whereas two and three epochs possess an apparent cumulative improvement.

5. Discussion

These two experiments indicate that AraBERTv2 is very effective on the Arabic news classification with moderate sequence lengths and enough steps in the fine-tuning process. The most practical setting that has been tested is the 128 tokens three epochs set. Such a combination is a balance between coverage (contextual and) and effective optimization.

Task wise the outcome of 128 tokens being better than 256 tokens has a meaning. In the headline of the Arabic news articles, the topic markers are frequently very informative. In the cases when the classification task is topical, the evidence can be already helpful and may be placed at the beginning of the article. It is possible to feed longer sequences and thus this increases costs and may not even improve performance. This finding is in line with more general practice in transformers, where more lengthy inputs need to warrant their representational and computational costs [10, 12].

The experiment on the epoch is also informative. The gradual increase in the number of epochs between one and three implies that the model had not entirely adjusted at the end of the initial run over the data. Accuracy and macro-F1 also improved concomitantly and thus the gains were not propagated by one dominant class. In local conference work, and also on practical work, this can be useful: it suggests that small extra training can be justified by the computational budget when feasible.

Another implication is on reproducibility. The Arabic NLP research is always prone to variation in models, preprocessing, data divisions, and data optimiser and it is hard to figure out why a certain set-up performs better than another. The current study offers a more simplistic experimental image that other researchers can replicate or expand because of the isolation of sequence length and epoch count at the expense of not changing other aspects of the pipeline.

The paper is still purposeful. It does not compare different Arabic backbones, does not optimize learning rate or warmup schedule, or tunings of variants of optimizers. However, the findings are practical since they give factual evidence regarding two of the most modified parameters in practice. The results may be employed as a benchmark in future studies of conferences that study how learning rate, data augmentation or cross domain transfer affect learning.

6. Practical Recommendations

In practical terms, the findings can be applied into a number of practical suggestions aimed at the Arabic news classification projects. First, researchers cannot just assume that the longest sequence length that could be afforded is necessarily the preferable one. Overall, in this work, 128 tokens performed better than 256 tokens despite the fact that the latter gives more texts to the model. In situation where the task is headline- and lead-driven classification of topics, a moderate length could be able to extract a majority of the useful signal and reduce memory and training time.

Second, the epoch experiment indicates performance on the table with short fine-tuning schedules. One epoch had respectable results, whereas at two and three epochs, the model had significantly better results. This is a key message when the local conference studies or classroom deployments: it can be demonstrated that the careful tuning of the training duration can provide huge gains without adding new architecture or more data collection.

Third, both the accuracy and the macro-F1 should be reported as mainstream to Arabic classification studies. Since Arabic datasets may differ in terms of their balance and difficulty of the labels, the adoption of a single measure can hide important information. The fact that accuracy and macro-F1 values are very similar in this study is promising since it shows that the improvements were not contained in a few easy classes.

Last but not least, there is reproducibility. Small Arabic NLP projects can be difficult to compare as preprocessing, sequence settings, splitting, and stopping criteria can be adjusted all at once. The current research reveals the usefulness of modifying individual factors. In case of follow-up work, one can use the same philosophy in the learning rate, the batch size, the warmup steps, or early stopping strategy.

7. Threats to Validity

A number of limitations must be mentioned. First, the experiments were done using one Arabic news dataset and one pretrained backbone. The trends identified might not be the same in other Arabic areas like dialectal social media, intent detection, or sentiment analysis.

Second, there are three values of sequence length and three values of epoch count, which are assessed in the current work. These options are feasible and reflective but a finer grid might indicate a more accurate optimum. Third, the findings are stated at a set learning rate and batch size; the connection between these parameters and the parameters under test was not investigated in the current paper.

8. Conclusion and Future Work

This paper provided a narrow empirical study of two useful hyperparameters of AraBERTv2-aided Arabic news classification maximum input sequence length and the amount of training epochs. The sequence-length experiment indicated that 128 tokens gave the optimal result in the experimented values whereas the epoch experiment indicated that the epochs with the highest accuracy and macro- F1 were three. The overall setup gave the highest accuracy and macro-F1 of 97.44 and 97.43.

The overall lesson is simple: sequence lengths should be modest, and the extension of fine-tuning should be moderate, and can make a significant positive difference to Arabic news classification, with no architectural alterations. This experimental framework could be applied in future work to learning-rate analysis, checkpoint selection, early stopping, cross-source transfer, and intra-Arabic model comparisons.

References

[1] M. Mashaabi, S. Al-Khalifa, and H. J. a. p. a. Al-Khalifa, "A survey of large language models for Arabic language and its dialects," 2024.

[2] Y. Matrane, F. Benabbou, N. J. J. o. K. S. U.-C. Sael, and I. Sciences, "A systematic literature review of Arabic dialect sentiment analysis," vol. 35, no. 6, p. 101570, 2023.

[3] N. Sengupta *et al.*, "Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models," 2023.

[4] A. Ghaddar, P. Langlais, M. Rezagholizadeh, and B. J. a. p. a. Chen, "On the importance of Data Scale in Pretraining Arabic Language Models," 2024.

[5] G. Bhatia, A. El Mekki, F. Alwajih, and M. Abdul-Mageed, "Swan and ArabicMTEB: Dialect-aware, Arabic-centric, cross-lingual, and cross-cultural embedding models and benchmarks," in *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025, pp. 4654-4670.

[6] F. J. a. p. a. Qarah, "Saudibert: A large language model pretrained on saudi dialect corpora," 2024.

[7] F. J. a. p. a. Qarah, "Egybert: A large language model pretrained on egyptian dialect corpora," 2024.

[8] A. S. J. T. J. o. S. Alammary, "Investigating the impact of pretraining corpora on the performance of Arabic BERT models," vol. 81, no. 1, p. 187, 2025.

[9] M. M. Hossain, M. S. Hossain, M. Safran, S. Alfarhood, M. Alfarhood, and M. F. J. I. A. Mridha, "A hybrid attention-based transformer model for Arabic news classification using text embedding and deep learning," vol. 12, pp. 198046-198066, 2024.

[10] L. H. Baniata and S. J. M. Kang, "Switching self-attention text classification model with innovative reverse positional encoding for right-to-left languages: a focus on Arabic dialects," vol. 12, no. 6, p. 865, 2024.

[11] M. M. S. Al Deen, M. Pielka, J. Hees, B. S. Abdou, and R. Sifa, "Improving natural language inference in Arabic using transformer models and linguistically informed pre-training," in *2023 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2023, pp. 318-322: IEEE.

[12] J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, and N. J. a. p. a. Smith, "Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping," 2020.

[13] M. Ahmed, S. Alfasly, B. Wen, J. Addeen, M. Ahmed, and Y. Liu, "AlclaM: Arabic dialect language model," in *Proceedings of the Second Arabic Natural Language Processing Conference*, 2024, pp. 153-159.

[14] A. Paran, S. Shohan, M. Hossain, J. Hossain, S. Ahsan, and M. M. Hoque, "Semanticcuetsync at AraFinNLP2024: classification of cross-dialect intent in the banking domain using transformers," in *Proceedings of the Second Arabic Natural Language Processing Conference*, 2024, pp. 422-427.

[15] Z. Ferhat *et al.*, "Functional Text Dimensions for Arabic Text Classification," in *Proceedings of The Second Arabic Natural Language Processing Conference*, 2024, pp. 352-360.

[16] Y. Al Hariri and I. A. Farha, "SMASH at AraFinNLP2024: Benchmarking Arabic BERT models on the intent detection," in *Proceedings of the Second Arabic Natural Language Processing Conference*, 2024, pp. 403-409.

[17] M. Alruily, A. Manaf Fazal, A. M. Mostafa, and M. J. A. S. Ezz, "Automated Arabic long-tweet classification using transfer learning with BERT," vol. 13, no. 6, p. 3482, 2023.

[18] N. Badri, F. Kboubi, A. J. A. T. o. A. Habacha Chaibi, and L.-R. L. I. Processing, "Abusive and hate speech classification in arabic text using pre-trained language models and data augmentation," vol. 23, no. 11, pp. 1-28, 2024.